DOCUMENT RESUME

ED 346 158 TM 018 510

AUTHOR Sungsook, Kim C.
TITLE The Dependability of Student Ratings of Instructors across Sections.
PUB DATE Apr 92
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Business Education; *College Students; Course Evaluation; Equations (Mathematics); Foreign Countries; Generalizability Theory; Higher Education; Mathematical Models; *Professors; Rating Scales; *Reliability; Student Attitudes; *Student Evaluation of Teacher Performance; *Teacher Effectiveness
IDENTIFIERS Hierarchical Analysis; Rater Reliability; South Korea

ABSTRACT

The consistency of ratings between two sections of a course taught by an instructor within a given semester was studied by examining sources of variation in rating instructor/course obtained through a college teacher evaluation system. Data were analyzed for five business school instructors at a Sogang University in Seoul (South Korea). Each professor taught more than two sections and had at least 10 ratings completed by students in each section. A 25-item rating form had been developed for this university. The generalizability design included a three-facet hierarchical design for section, student, and item. Results indicate that teacher performance in a given course is very consistent from section to section. If an instructor's performance is being assessed, generalizability is enhanced by combining ratings from more than one class, mainly because this increases the number of students sampled. There is a possible problem in validity associated with the use of attribute items. Averaging the responses to several attribute items may produce an index with high generalizability, but the different attributes may vary in importance to students in different courses or departments. There are five tables of study data and a 14-item list of references. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
SUNGSOOK KIM C.
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED346158

THE DEPENDABILITY OF STUDENT RATINGS OF INSTRUCTORS ACROSS SECTIONS

Sungsook Kim C.
Sogang University, Seoul, Korea

Paper presented at the annual meeting
of the American Educational Research Association

San Francisco, April, 1992

TM018510

BEST COPY AVAILABLE

THE DEPENDABILITY OF STUDENT RATINGS OF INSTRUCTORS ACROSS SECTIONS

I. Introduction

Much of the research on teaching is directed toward the goal of ultimately identifying cause and effect relationships between classroom process variables, such as teacher behaviors, instructing methods, student product variables, in order to improve educational outcomes. Determination of whether causal relationships exist between process and outcomes of instruction, so called 'teaching effectiveness', is dependent on experimental manipulation of variables which have been previously identified, often through observational studies in the classroom. The instructing process evaluated by students now has become an issue even in Korea where having such a saying, "you must not step on the shadow of a teacher".

However, when researchers or college administrators implement student ratings of college instruction, one problem that must be faced is the assessment of the reliability of a measure used in the rating system. Measures of instructor (or teacher) performance (as rated by students) contain a number of potential sources of error associated with courses, with semesters, with sections, with students, with items of instrument, and with their combinations. Several studies have applied generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) to estimate generalizability coefficient enables the researcher to analyze the influence of multiple sources of variation in the ratings of the instructors. Crook and Kane (1976; 1981) estimated variance components for the students and item effects and found that the variability of student responses was substantially larger than the variability due to item selection. A study by Smith (1979) examined the variance associated with different occasions on which a given course was taught by an instructor. Using data from three consecutive semesters, he found the variance components associated with occasions to be very small, hence concluded that

perceived instructor performance in a particular course was very stable from occasion to occasion. More studies (Gillmore,et al.,1978;Thompson & Melancon,1987) confirmed the earlier findings that student effects were more important than item effects, but produced somewhat inconsistent results with regard to the relative magnitudes of the teacher and course effects. On the other hand, Smith (1986) estimated variance components for instructor and course items that the teacher effect was much larger than the course (within teacher) effect. However, it is evident from these studies that ratings of individual instructor often vary quite substantially from course to course.

The present study examined the section effect, that is, the consistency of ratings between two sections of a course taught by an instructor within a given semester. More importantly, since the component behaviors could be rated more objectively than the global domain of behaviors (Thompson et al.,1987), the generalizability of instructor ratings obtained used specific attribute items. Once the magnitude of each source of variation had been identified, their effects would be compared global(very general) instructor items with attribute items (dealing with different aspects of the instructor's performance).

Therefore, the purpose of the study was to examine the sources of variation in rating instructor/course which were attained through an evaluation system in a college. Specifically, this study examined the magnitude of variation due to section (within teacher), students (within section), items and the interaction among these factors for four global items and twelve attribute items. Second emphasis of the study was to present various combinations of numbers of items, students, and sections such that a satisfactory level of generalizability was attained and compared across different types of items.

II. Methodology

Data

The data for this study were based on the course and instructor evaluation conducted in the spring semester of 1991 for the Business

BEST COPY AVAILABLE

Department of Sogang University in Seoul, Korea. The sample consisted of five instrutors who taught more than two sections of the same course,(such as, business statistics,business data management,etc.) and had at least 10 adequately completed rating forms randomly chosen from more than 30 students for each section. Students in each section rated their instructor in the final class of the course.

The course/instructor evaluation form used in the study was based on an extensive review of the indicators of process-product relationship based on the issue of the teaching effectivenss in the college. The form developed in 1990 was composed of 25 items, 16 of the 25 items on the form were relavant to this study : four very general (global) items and 12 more specific items examing a reasonably broad sample of instructor attributes and course organization, that is, clarity of presentation of course material, ability to answer questions, vocal delivery and speech clarity, use of the blackboard, attitude toward students, enthusiasm in teaching the course, interaction with students, appropriateness of exams and assignments, preparation and structure of lecture, and an evaluation standard for the course,etc.

Design

The generalizability design used to address the questions included a three-facet hierarchical design, three facets were section and student and item. The design was a nested analysis of variance design ((s:c:t) x i) with section or class (c) nested within instructor or teacher (t), students (s) nested within classes, and items (i) crossed with the other three factors. The conditions of each facet can be defined as a sample of a complete set of conditions(i.e., fixed effect) or as the infinite set of conditions(i.e.,random effect). For the design, it was assumed that instructors were a fixed effect because 5 instructors were chosen among 12 instructors who had taught two or more sections of the same course. However, sections were treated as a random effect because the sections were randomly selected from two or more sections of each course. Students were also considered to be a random effect since the students of each section were representative to the second or the third year in the business

school. However, items on the form were treated as a fixed effect because the four global items and the twelve specific items had been deliverately selected by the researcher.

Analysis

Because students were nested within sections and the sections were nested within teachers, there were only seven varinace components in this analysis. This occurred because the nested design confounded the variation attributed to sections with the interaction teacher and section, and confounded the student-section interaction with the three way interaction among students, sections, and teachers. Therefore, the rating of the teachers' use of likert scale, for a particular instructor(t) for each section(c) rated by a particular student(s) on a particular item, can be decomposed into eight components as follows:

$X_{scti} = \mu$	grand mean
$+ (\mu_t - \mu)$	instructor effect
$+ (\mu_{c:t} - \mu_t)$	section:instructor effect
$+ (\mu_{s:c:t} - \mu_{c:t})$	student:section:instructor
$+ (\mu_i - \mu)$	item effect
$+ (\mu_{ti} - \mu_t - \mu_i + \mu)$	instructor x item
$+ (\mu_{ci:t} - \mu_{c:t} - \mu_{ti} + \mu_t)$	item x(section:instructor)
$+ (X - \mu_{s:c:t} - \mu_{ci:t} + \mu_{c:t})$	residual

$$\sigma_x^2 = \sigma_t^2 + \sigma_{c:t}^2 + \sigma_i^2 + \sigma_{ti}^2 + \sigma_{ci:t}^2 + \sigma_{s:c:t,si:c:t}^2$$

In other words, the variance of the ratings can be partitioned into independent sources of variation due to differences between instructors, sections, students, items, their interactions, and the residual. In the notation for those components, the colon(:) implied 'nested within', while two consecutive subscripts implied crossing of the effect.

Separate analyses were carried out for four global items and twelve attribute items. From each ANOVA procedure, an estimate of the variance components corresponding to each factor and to each

BEST COPY AVAILABLE

interaction between factors was calculated from the mean squares. The estimated variance components were then compared for relative magnitudes and estimated by modifying the number and contribution of facets. Table 1 shows how to estimate each variance components from the analysis of variance by setting the expected mean squares.

Table 1. Three-facet, Partially nested I x (S:C:T) design
Equation for estimates of variance components
(Teacher:fixed, Class:random, Student:random, Item:fixed)

SV	df	σ^2
t	t - 1	(MS_t - $MS_{c:t}$ - $MS_{ci:t}$) / csi
c:t	t(c - 1)	($MS_{c:t}$ - $MS_{s:c:t}$ - $MS_{si:c:t}$)/ si
s:c:t	tc(s - 1)	($MS_{s:c:t}$ - $MS_{si:c:t}$) / i
i	i - 1	(MS_i - MS_{ti}) / tcs
ti	(t - 1)(i - 1)	(MS_{ti} - $MS_{ci:t}$) / sc
ci:t	t(c - 1)(i - 1)	($MS_{si:t}$ - $MS_{si:c:t}$) / s
si:c:t	tc(s - 1)(i - 1)	$MS_{si:c:t}$

A generalizability coefficient is analagous to a classical reliability estimate except that distinct sources of measurement error are recognized and accounted for by the generalizable universe score. The instructor was the object of measurement in this study, the variance component for instructors(σ_t^2) represented the universe score variance. The error variances included the variation related to the object of rating, that is, the variance of section within instructor($\sigma_{c:t}^2$), interaction between teacher and item(σ_{ti}^2), interaction among item and section nested within instructor($\sigma_{ci:t}^2$), and residuals($\sigma_{s:c:t,si:c:t}^2$). A generalizability study can generate several coefficients, each corresponding to a different universe of conditions. The generalizability coefficient specified from the design used in this study can be expressed as follows:
N indicated the conditions(numbers) of each facet, that is, section, student, and item.

$$\rho^2 = \frac{\sigma_t^2}{\sigma_t^2 + \frac{\sigma_{c:t}^2}{N_c} + \frac{\sigma_{ti}^2}{N_i} + \frac{\sigma_{ci:t}^2}{N_{ci}} + \frac{\sigma_{s:c:t,si:c:t}^2}{N_{sci}}}$$

The estimated variance components were then compared for relative magnitudes, and the generalizability coefficients of interest in which generalization of instructor's performance was over section, students, and items were obtained. The coefficient may be interpreted as the expected correlation between the ratings by two randomly chosen sets of students, in two randomly chosen section of a given course, on two sets of items(four global and twelve attribute items). The data were analyzed with SPSS/PC+ MANOVA four factor nested design and the GENOVA computer package developed by Brennan (1983).

III. Results

The analysis of the ratings of course and instructor revealed the global and the attribute items were generalizable. Table 2 presented the results of ANOVA and the estimated variance components for the global items. For the attribute items, six items related to course evaluation and the rest six related to instructor evaluation were analyzed separately.

Insert Table 2 here.

For the global items, for a combination of one section and one student and one item, the instructor variance component(.285) was relatively large. However, while the variance component of items(.003) was very small, the variance component of instructor-item interaction (.114) was relatively large. The variance component of section within instructor(.063) was also relatively small, but that of interaction between item and section within instructor was negative. Variance

components must be non-negative, but the estimates obtained by the procedures given can be negative because of sampling errors. This substituted to zero and carried this zero forward as the estimate of the component when it entered any equation (Cronbach, et al. 1972).

The findings indicated that the instructors rated differently across cources, but they who taught the same course acted similarly. Also, the small variance component of items pointed the global items were rated very similarly across the different types of questions. In order to compare the relative magnitude of the estimated variance components for each facet and interactions, the estimates of universe score variance and error variance for both groups of items were summarized in Table 3.

Insert Table 3 here.

For the course-related attribute items, the magnitude of variance componenet related to instructors(.065) was similar to that related to items(.060). The section within instructor variance(.046) and the interaction between item and instructor (.047) were about same and the variance component of section-item interaction was zero. However, for the instructor-related attribute items, the variance component of items (.174) was about four times larger than that of instructor(.044). The variance of interaction between section and item was quite small(.009), and the variance related to section within instructor and that related to instructor-item interaction were zero. The residual variances of the attribute items were relatively large (.80, .88, respectively). The findings indicated that the section and the item-section interaction variance components were relatively small or zero, therefore, the variance related to sections had little effect on decreasing the magnitude of error variance.

Since the generalizability coefficient was calculated based on a combination of one section, one student and one item, it can be improved by modifying the number of section, student, or item. The findings showed the effect of sections was less than that of items, therefore, the optimal study for improving the generalizability coefficient of the score would be based on having different

combinations of fewer sections and more items. Table 4 and Table 5 presented the result of estimated generalizability coefficient for different combinations of number of sections, students, and items for global items and course-related attribute items. From the contribution of each error variance component to the universe variance, there was a table which included the possible combination of students, items, and sections needed to attain the specified acceptable level of generalizability, .80. For both groups of items, the generalizability coefficient increased considerably as the number of students increased. However, the results for increasing the number of global items and attribute items differed considerably. There was little gain in generalizability from using more than one global item, but substancial gain from using additional attribute items.

Insert Table 4 and Table 5 here.

For the global items, the level of generalizability of at least .80 was obtained at least with the combination of 2 sections, 2 students and 10 items or 2 sections, 5 students and 5 items. For the attribute items, the item effect had a relatively strong effect on the generalizability of ratings. The combination of 4 sections, 10 students, and 20 items needed to attain the accepable level. Increasing the number of items produced a better generalizability coefficient more efficiently than increasing the levels of other facets. Therefore, although the combination of different number of each condition was applied, having the combination of fewer section and more students and items produced better generalizability (e.g. 3 sections, 5 global and 20 attribute items and 10 students).

IV. Discussion

The results of the study indicated that teacher performance in a given course was very consistent from section to section. This findings suggested that, if an instructor's performance in a given

BEST COPY AVAILABLE

course is being assessed, generalizability was enhanced by combining ratings from more than one class, but mainly because this effectively increased the number of students sampled. Also, the global items were measured more generalizably than the attribute items. Approximately five attribute items were needed to equal generalizability from one global item. These 'item specificity' effects are substancially stronger than those observed by other studies (Smith, 1979). A plausable reason for this difference is that this instrument(evaluation form) used more a diverse sample of attribute items.

Comparing global items with attribute items, the magnitude of the instructor variance of global items was larger than that of attribute items. Therefore, based on one section, one student, and one item, the generalizability coefficient of the global items was much larger than that of the attribute items (.20,.07,respectively). However, analyzing the variance related to the error variance enabled researcher to estimate each influences of multiple sources of variation.

The study examined possible combinations of the conditions of each facet in order to determine the number of items, students and sections that are needed to obtain an acceptable level of generalizability for a measure. For both global and attribute items, the generalizability coefficients increased considerably as the number of students increased, but the number of items had a considerable effect only for attribute items. Also,because the section and item-section interaction variance components were zero or small, increasing the number of class(section) had approximately the same effect as adding a corresponding number of students to a single class. The findings suggested that if an instructor's performance in a given course was being assessed, generalization was enhanced by combining ratings from more than one section, but mainly because this effectively increased the number of students sampled. In addition, there was little gain in generalizability from using more than one global item, but substantial gain from using additional attribute items. This indicated the generalizability of ratings was substantially influenced by item specificity.

However, if students ratings are used to provide an overall index of instructor acceptability to students, as is frequently the case

where ratings are use for administrative purposes, there is a possible problem in validity associated with the use of attribute items. Averaging the responses to several attribute items may indeed produce an index with high generalizability, but the different attributes may vary considerably in importance to students in different courses or departments.

Table 2. Results of ANOVA and Estimates of variance component for Global items

SV	df	MS	F	σ^2
t	4	26.81	7.66 *	.285
c:t	5	3.5		.063
i	3	3.09	1.12	.0033
ti	12	2.76		.114
ci:t	15	.48		.000
s:c:t,si:c:t	90,270	.97		.97

Table 3. Estimates of Universe score variance and Error variance for Global and Attribute items

	Universe variance σ_t^2	Error variance $\sigma_{c:t}^2$	σ_{ti}^2	$\sigma_{ci:t}^2$	$\sigma_{s:c:t,si:c:t}^2$
Global	.285 (20%)	.063 (4.4%)	.114 (7.6%)	.000 (.00%)	.97 (67%)
Course attribute	.065 (6.8%)	.046 (4.8%)	.047 (4.9%)	.000 (.00%)	.80 (83.5%)
Instructor attribute	.0435 (4.6%)	.000 (.00%)	.009 (1.0%)	.009 (1.0%)	.88 (93.4%)

Table 4. Estimates of generalizability coefficient using various combinations of classes(sections), students, and items for global items

item	classes (sections) 1				2				4				8			
	students 1	2	5	10	1	2	5	10	1	2	5	10	1	2	5	10
1	.20	.30	.43	.51	.31	.42	.54	.59	.43	.45	.61	.65	.54	.61	.66	.68
2	.32	.44	.57	.63	.46	.49	.67	.71	.50	.68	.70	.77	.69	.75	.78	.80
5	.50	.61	.69	.73	.65	.73	.79	.82	.76	.82	.85	.87	.84	.87	.89	.90
10	.62	.70	.75	.77	.76	.81	.84	.86	.85	.88	.90	.91	.90	.92	.92	.93
20	.71	.75	.78	.80	.82	.85	.87	.88	.89	.90	.92	.93	.93	.94	.95	.96

Table 5. Estimates of generalizability coefficient using various combinations of classes(sections), students, and items for course-related attribute items

item	classes (sections) 1				2				4				8			
	students 1	2	5	10	1	2	5	10	1	2	5	10	1	2	5	10
1	.07	.12	.20	.27	.12	.19	.30	.37	.20	.28	.39	.44	.30	.38	.47	.50
2	.12	.19	.30	.37	.21	.31	.43	.49	.32	.42	.52	.56	.44	.53	.61	.64
5	.23	.32	.43	.48	.37	.47	.57	.62	.50	.59	.66	.69	.64	.70	.75	.77
10	.33	.42	.49	.53	.49	.58	.64	.67	.61	.67	.72	.74	.74	.78	.81	.82
20	.42	.49	.53	.55	.59	.65	.69	.70	.69	.73	.76	.77	.81	.83	.85	.86

References

Brennan, R. (1983). Elements of Generalizability Theory. Iowa City, IA: The American College Testing Program.

Choi, S.K., & Loyd, B.H. (1987). Sources of variation in the recording of teacher behaviors in the classroom. Paper presented at the annual meeting of the Eastern Educational Research Association, Boston.

Choi, S.K., & Loyd, B.H. (1988). An analysis of sources of variance in an observational system. Paper presented at the annaul meeting of the American Educational Research Association, New Orleans.

Cronbach, L.J., Glesser, G.C., Nanda, H., & Rajaratnam, N. (1972). The generalizability of behavioral measurement : The theory of generalizability for scores and profiles. New York : John Wiley.

Erlich, O., & Shavelson, R. J. (1978). The search for correlations between measures of teacher behavior and student achievement : Measurement problem, conceptualization problem, or both ? , Journal of Educational Measurement, 15, 77-89.

Gillmore, G.M., Kane, M.T., & Naccarato, R.W. (1978). The generalizability student ratings of insruction : Estimation of the teacher and course components. Journal of Educational Measurement, 15, 1-13.

Lomax, R.G. (1982). An application of generalizability theory to observational research. Journal of Experimental Education, 51, 22-30.

Lord, F.M. & Novick, M.R. (1968). Statistical Theories and Mental Test Scores. Reading MA : Addison-Wesley.

Shavelson, R., Webb, N., & Burstein, L. (1986). Measurement of teaching. In Gage, N. L. (3rd Ed.). Handbook of Research on Teaching. Chicago, IL: Rand McNally.

Smith, P. L. (1978). Sampling errors of variance components in small sample multifacet generalizability studies. Journal of Educational Statistics, 3, 319-346.

Smith, P. L. (1979). The stability of teacher performance in the same course over time. Research in Higher Education, 11, 153-165.

Smith, P., Waller, M., & Waller, S. (1982). Generalizable obsevation of the teaching process. Educational and Psychological Measurement, 42, 467-478.

Thompson, B. & Melancon, J.G. (1987). Measurement characteristics of the group embedded figure test. Educational and Psychological

Measurement, 47, 765-772.

Willson, V., Mangano, N. & Rupley, W. (1986). Sources of variation that affect the reliability of reading classroom observation measure. Paper presented at the annual meeting of the American Educational Research Association.